ABSTRACT

A system for and method of allocating a resource to a service request based on application of a persistence policy is described. In one embodiment, upon or after allocation of a resource to a resource request, an entry representing the allocation is made in a data structure using a first index derived from information relating to the resource request if such is available. An entry representing the allocation is also made in the data structure using a second index derived from information relating to the resource request. When a resource request is received, the data structure is accessed using the first index if such is available. If an entry corresponding to the first index is available, the resource corresponding to the entry is allocated to the request. If the first index or an entry corresponding to the first index is unavailable, the data structure is accessed using the second index. If an entry corresponding to the second index is available, the resource corresponding to the entry is allocated to the request. If an entry corresponding to the second index is unavailable, a resource is allocated to the request based on application of a load balancing policy or other persistence policy. One or more entries are then made in the table corresponding to the new allocation.